Robust Transformer Neural Network for Computer Vision Applications

Fazlur Rahman Bin Karim[1], Dimah Dera[2]

1Electrical and Computer Engineering, University of Texas Rio Grande Valley, Edinburg, TX.
2Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY.

INTRODUCTION

Computer Vision Models are susceptible to Noises and Adversarial Attacks and hereby hinder the widespread application in crucial domains such as medical diagnosis of tumors and object detection in self-driving vehicles. With this research, we aim to develop a **Robust Framework for the Vision Transformer (ViT)** model using the Bayesian Framework that simultaneously predicts the output in the Image Classification task and quantifies uncertainty in the prediction.

PURPOSE AND HYPOTHESIS

The Self-Attention structure of Transformer has emerged as an alternative to **Convolutional Neural Networks (CNNs)** in numerous computer vision applications due to the success garnered in **Natural Language Processing (NLP).** Although ViT models have demonstrated outstanding predictive performance, they lack sufficient **uncertainty quantification** and are prone to overconfident predictions like other deep learning techniques when faced with perturbation. That makes Vit models unreliable for critical domains, i.e., healthcare and transportation [1].
Our contribution lies in :

- ❑Developing a **Robust Vision Transformer** Models utilizing Bayesian Framework.
- ❑Propagating **Mean** and **Covariance** through Different Layers and non-linear functions of ViT to quantify uncertainty .
- ❑**Mean** delineates **prediction of output** where covariance matrix depicts the **uncertainty estimation** in the predicted decision.
- ❑Proving superior robustness of the developed model by comparing with state-of-the-art model for benchmark datasets (MNIST, Fashion-MNIST, and CIFAR-10.)

METHODOLOGY

Data Preprocessing

The proposed model is validated on numerous benchmark image datasets (MNIST, Fashion-MNIST etc.) for Image Classification task.

The Bayesian Vision Transformer only takes sequence of token embeddings as an input. So, we need to create small fixed size non overlapping patches from input images and convert them into patch embeddings. Then resultant sequence of vectors are given to Bayesian Transformer Encoder .

- At first, Sequence of flattened 2D image patches $\{\mathbf{x_p} \in \mathbf{R^{N \times (p^2.c)}}\}$ are created from input image.
- After adding 1D Positional Embedding to the patch embeddings, the input sequence are given as an input to the proposed Bayesian Vision Transformer Model.

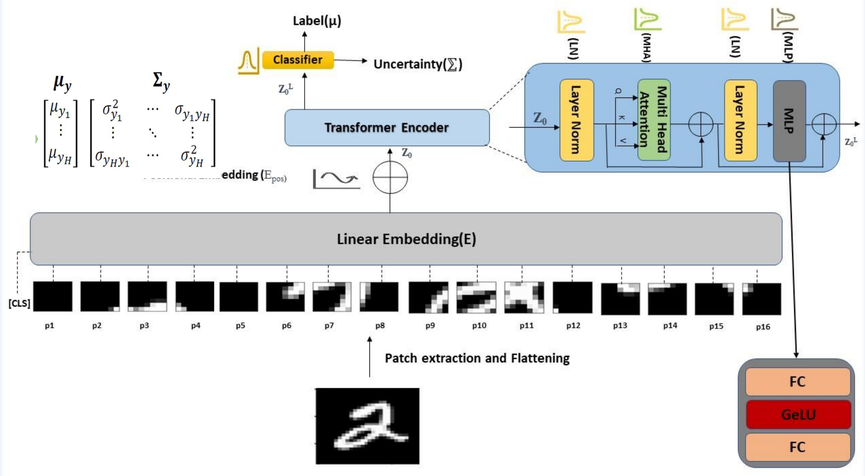

Fig 1: Illustration of the proposed Image Classification Technique based on Robust Bayesian Transformer Neural Network using MNIST dataset. (a) The input sequence comprised of 2D flattened patch embeddings $(\mathbf{x_p} \in \mathbf{R}^{N \times (p^2.c)})$ created from non-overlapping fixed sized patches enters Robust Transformer Encoder, (b) The Bayesian ViT model extracts features of the input sequence from the embedding matrix and processes these features through the propagation of the variational moments in different layers. (c) The internal structure of the encoder layer extracts relevant information between different tokens of the input image. (d) An expanded view of the Bayesian Transformer Encoder demonstrates the propagation of Variational Moments through its layers including the Multi-Head Attention (MHA), Layer Normalization, Multi-Layer Perceptron (MLP) along with Final Classifier. (e) The output fully connected layer classifies the predicted label and provides the uncertainty associated with the prediction through the covariance matrix.

Model Implementation

In our Bayesian Transformer Neural Network

- Input from each time step of the given input sequence is fed to the Transformer Encoder unit having network parameters (weights and biases) with a prior distribution.
- As illustrated in Figure 1, the objective is to obtain image classification along with uncertainty associated with the output .

RESULTS AND DISCUSSIONS

Table I: Test accuracy (in %) using Robust Bayesian Vision Transformer and Deterministic Vision Transformer for Image Classification Task using MNIST and Fashion-MNIST datasets under Gaussian noise,, FGSM and PGD adversarial attacks.

MNIST				Fashion-MNIST			
Robust Vision Transformer			Deterministic Vision Transformer	Robust Vision Transformer			Deterministic Vision Transformer
Noise Level				Noise Level			
No Noise		90.08	88.1	No Noise		82.44	79.9
Gaussian Noise	0.05	90.01	85.57	Gaussian Noise	0.05	81.60	79.2
	0.1	89.53	84.57		0.1	75.40	72.4
	0.2	86.50	78.22		0.2	52.20	47.7
FGSM	0.001	89.43	85.7	FGSM	0.001	81.10	79.23
	0.005	89.43	84.88		0.005	77.97	76.05
	0.01	89.35	83.51		0.01	70.50	69.4
	0.05	87.01	56.52		0.05	51.50	28.9
PGD	0.001	89.59	86.07	PGD	0.001	82.00	79.01
	0.005	89.53	85.12		0.005	80.30	77.2
	0.01	89.34	85.0		0.01	78.60	73.2
	0.05	88.59	80.92		0.05	68.00	45.9

✓ Higher accuracy with increased noise levels which justifies its '**robustness**'

CONCLUSION

The Robust Vision Transformer model demonstrates the **'robustness'** under high noise levels or stronger adversarial attacks. Such behavior can be used by the model to assess its own performance in high stake applications.

BIBLIOGRAPHY

[1] S. Yadav, S. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis,". J Big Data 6, 113 (2019).
[2] A. Dosovitskiy *et al.,* "An Image is Worth 16x16 Words: Trans formers for Image Recognition at Scale," in Proc. 9th Int. Conf. Learn. Representations (ICLR 2021), Austria, May 3-7, 2021.
[3] D. Dera, N. C. Bouaynaya, G. Rasool, R. Shterenberg and H. M. Fathallah-Shaykh, "PremiUm-CNN: Propagating Uncertainty Towards Robust Convolutional Neural Networks," IEEE Trans. Signal Process., vol. 69, pp. 4669-4684, 2021.

Acknowledgment

The work was supported by the **National Science Foundation CRII-2153413** Award.